Displays 70 (2021) 102102

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Multi-view stereo in the Deep Learning Era: A comprehensive review

Xiang Wang [a,1], Chen Wang [a,1], Bing Liu [b], Xiaoqing Zhou [a], Liang Zhang [a], Jin Zheng [c], Xiao Bai [a,*]

[a] School of Computer Science and Engineering, State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University, China
[b] Chinese Academy of Ordnance Sciences, Beijing, China
[c] School of Computer Science and Engineering, Beihang University, China

ARTICLE INFO

Keywords:
Multi-view Stereo
3D Reconstruction
Plane Sweep
Volumetric Representation
Deep Learning

ABSTRACT

Multi-view stereo infers the 3D geometry from a set of images captured from several known positions and viewpoints. It is one of the most important components of 3D reconstruction. Recently, deep learning has been increasingly used to solve several 3D vision problems due to the predominating performance, including the multi-view stereo problem. This paper presents a comprehensive review, covering recent deep learning methods for multi-view stereo. These methods are mainly categorized into depth map based and volumetric based methods according to the 3D representation form, and representative methods are reviewed in detail. Specifically, the plane sweep based methods leveraging depth maps are presented following the stage of approaches, i.e. feature extraction, cost volume construction, cost volume regularization, depth map regression and post-processing. This review also summarizes several widely used datasets and their corresponding metrics for evaluation. Finally, several insightful observations and challenges are put forward enlightening future research directions.

1. Introduction

3D reconstruction in various environments is one of the central tasks in 3D computer vision. It is of greater importance in artificial intelligence and has a wide spectrum of applications in automated driving, virtual reality/augmented reality [1,2], artificial intelligence robots, and other areas. With the increase and development of the 3D acquisition technology, depth sensors, such as LiDARs, are becoming reliable, lightweight and cheap, making them widely equipped with autonomous vehicles, ground robotics and even smart cellphones. However, depth maps captured by sensors are either sparse, losing abundant details (e.g., point clouds captured by LiDARs), or limited in a certain depth range, hindering the usage in outdoor scenes (e.g., Structured light or time-of-flight cameras). The demand for dense and detailed 3D reconstruction in a variety of scenes promotes the development of 3D reconstruction approaches from a series of images, which contain more texture and lighting information that is beneficial for reconstructing delicate models [3–5].

A general pipeline for image-based 3D reconstruction includes: 1) feature extraction and matching across multiple images for correspondence search; 2) image registration and triangulation for camera extrinsic parameter estimation and sparse reconstruction (also known as Structure from Motion); 3) dense 3D reconstruction from images using given intrinsics and estimated extrinsics. The dense 3D reconstruction stage given the calibrated cameras is vital as it directly related to the final quality of 3D reconstruction.

Multi-view stereo (MVS) is a fundamental component for 3D reconstruction, which estimated the dense representation of the 3D models from multiple overlapping images, utilizing stereo correspondence as the main cue [6,7]. Multi-view stereo approaches are commonly categorized according to the scene representation into depth map-, point cloud-, mesh-, and volumetric-based methods. The depth map representation presents the 3D geometry in a 2.5D form for each observation view. And the 3D reconstruction can be recovered by leveraging a 3D fusion technique to fuse the depth maps into a single coherent model. Point cloud representation depicts the scene in a sparse form, and it is usually obtained by back projecting image pixels along the viewing rays, on which the position of the points are determined by the depth maps. Volumetric representation describes the geometry using a regularly sampled 3D grids. A discrete occupancy function or a

* Corresponding author.
E-mail address: baixiao@buaa.edu.cn (X. Bai).
[1] Equal contribution.

https://doi.org/10.1016/j.displa.2021.102102
Received 30 July 2021; Received in revised form 21 September 2021; Accepted 24 September 2021
Available online 9 October 2021

continuous function encoding the distance to the closest surface is the usual form of volumetric representation. The triangular mesh-based surface is another coherent representation for 3D model, which can also be obtained via fusing multiple 2.5D depth maps.

Recently, deep neural networks (DNNs) have achieved outstanding predictive performance and have become an indispensable tool in a wide range of computer vision applications, e.g., person re-identification [8], image alignment [9], face alignment [10], object recognition [11,12] and stereo matching [13]. Their convincing performance on informative feature extraction and aggregation also triggers the interest to improve the multi-view stereo task. The rich representation of images helps in dealing with the difficulty of matching ambiguity caused by occlusion, varying lighting conditions or textureless regions. Thanks to the large-scale 3D scene reconstruction datasets, deep learning based multo-view stereo methods have made impressive performance improvement over traditional methods.

Multi-view stereo methods based on deep learning can also be divided according to the scene representation. However, currently only two major research directions exist, namely, depth map based methods and volumetric based methods. As the name suggests, depth map based networks predict the 2.5D depth map for each view, utilizing the geometric information among overlapping views. All depth maps are later fused into other coherent 3D representations using some depth fusion and filtering methods as the post processing step. Volumetric based networks, in contrast, directly predict the occupied position in the 3D voxelized space from the input image set as the globally coherent scene representation. While the computational cost of depth map based methods depends on the number and resolution of input images, that of volumetric based methods is directly related the scale of reconstructed scenes, hampering the volumetric based methods from reconstructing large-scale outdoor scenes. Thus, most multi-view stereo networks produce the depth maps for each image observation and constructed a single 3D model indirectly.

In the past years, works in the area of multi-view stereo methods based on deep learning has developed rapidly, especially after the work of MVSNet [14] which leveraged the plane sweep based cost volume formulation to predict the depth maps. In this review we focuses on the recent advanced of deep learning methods for multi-view stereo, including both depth map based and volumetric based approaches and mainly focusing on the former. We introduce some important details of these works. Specifically, we present the depth map based methods with four major stages, i.e., feature extraction, cost volume construction, cost volume regularization, depth regression and post processing. Among them, the cost volume construction and regularization are the main research focus, since they significantly impact the accuracy of depth maps, and take the most computational resources. Also few works are based on volumetric representation, there is still a trend to effectively and efficiently construct a coherent 3D volumetric model using networks. Thus we mentioned several representative works in this review. This paper also introduces several representative datasets and metrics for performance evaluation, and compares performance of several works on the most mainstream datasets.

The major contributions of this article can be summarized as follows:

- To the best of our knowledge, this is the first review that covers the recent advances of deep learning based multi-view stereo methods, including both the depth map based and volumetric based ones.
- The depth map based methods are analyzed in detail, presenting the main focus of recent works.
- This paper summarizes the performance of most methods on several mainstream datasets.

In this paper, we provide a review of the deep learning methods for multi-view stereo, which is organized as follows: Section 2 provides a detailed survey of the depth map based methods. Section 3 provides a survey of the volumetric based methods. Section 4 introduces several widely used datasets for the training and evaluation of multi-view stereo networks, and summarizes the performance of most methods on two major datasets. Section 5 concludes the paper and gives a brief discussion on the future research directions in this topic.

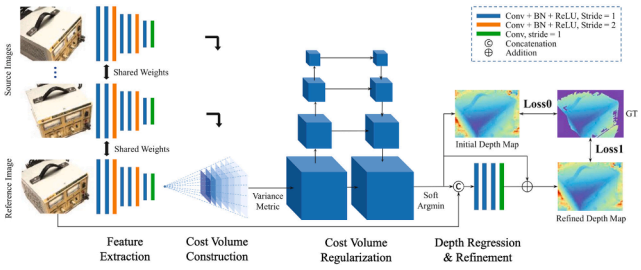

Fig. 1. The overall Structure of MVSNet [14].

2. Depth Map Based Methods

Given the calibrated camera parameters of all views, the 3D models can be uniquely projected into the color images and depth maps. Depth map based multi-view stereo methods thus aims to recover the 3D models by predicting the depth map of each view. Regarding multi-view images as the combination of several stereo image pairs, classical binocular stereo matching techniques such as cost computation and aggregation can be used in multi-view depth map computation. The cost volume in binocular stereo matching contains the matching cost for all depth hypotheses. Multi-view stereo generates the cost volume via the so-called "plane sweep" method, which "sweeps" several frontal-parallel virtual hypothetical planes in the 3D space using the homography matrices computed by camera parameters, and computes the photometric consistency as a (proxy) measure of depth likelihood. The depth estimate can thus be derived from the cost volume.

Specifically, the plane sweep is conducted using the differentiable warping operation. Given a pair of images with known intrinsics $\mathbf{K}_r, \mathbf{K}_s$ and extrinsics $\mathbf{T}_r, \mathbf{T}_s$ ($\mathbf{T}_i = [\mathbf{R}_i, \mathbf{t}_i]$ including rotations and translations, $i = r$ *or* s) of the reference and source camera respectively, the hypothesis depth d in the reference frame and the principle axis n of the reference camera, a homography matrix $\mathbf{H}_s$ can be computed that determines the coordinate mapping in the warping operation from the pixel location u_r in the reference view to $u_s \sim \mathbf{H}_s(d)u_r$ in the source view:

$$\mathbf{H}_s(d) = \mathbf{K}_s\mathbf{R}_s(\mathbf{I} - \frac{(\mathbf{t}_r - \mathbf{t}_s)\mathbf{n}^T}{d})\mathbf{R}_r^T\mathbf{K}_r^{-1}. \quad (1)$$

The differentiable warping process relates the pixel pairs corresponding to the same 3D points and makes cost volume construction in networks possible.

A typical plane sweep based multi-view stereo framework consists of the following procedures: feature extraction, cost volume construction, cost volume regularization, depth regression and post-processing. MVSNet [14] presented a standard end-to-end depth estimation network for multi-view stereo, whose full structure is shown in Fig. 1. The deep features was extracted from the multiple input images using an eight-layer CNN whose weights were shared across different views, which contain rich contexts for dense matching. Given several hypothesis planes determined by a pool of depth hypotheses (uniformly sampled within a depth range), the cost volume for the reference image could be constructed based on the differentiable homography and a suitable cost metric. Here the element-wise variances of warped features from the reference images and all source images were utilized as the cost metric, which resulted in a 4D cost volume that aggregated information from all views. To suppress noise due to factors violating the consistency assumption (e.g., non-Lambertian surfaces or object occlusions), the cost volume should be regularized with smoothness constraints. A multi-scale 3D U-Net was introduced to regularize the cost volume, which

aggregating contexts from a large reception field with relatively low computational cost. The regularization module finally output a 3D cost volume (i.e. 1 channel in feature dimension, denoting as the latent probability volume), which passed through a softmax operation along depth dimension to form a probability volume. To estimate the final depth for the reference image, a direct operation was to choose the most likely depth hypothesis from the probability volume (i.e. argmax). Nevertheless, such winner-take-all operation was neither differentiable nor able to give a sub-pixel estimation. Instead, the so-called soft argmin operation [15] was used to regress the depth. Concretely, the expectation value of depth was computed as the output depth, i.e.

$$d = \sum_{d'=d_{min}}^{d_{max}} d' \times p(d') \tag{2}$$

where $p(d')$ was the probability of the hypothesis d' and $[d_{min}, d_{max}]$ defined the range of uniformly sampled depth hypotheses. Note that the large reception field involved in cost volume regularization could oversmooth the reconstruction boundaries, the depth map output needed to be refined under the guidance of well-textured reference image. Taking the reference image and the initial depth map as input, an additional convolutional network was applied to produce a residual depth map, which recovered depth details and was added back to the initial depth map to produce a refined depth map. To train MVSNet, a supervised L_1 loss was applied on both the initial depth map and the refined one. As the depth map was predicted per view, errors in the prediction would cause inconsistency when merging multiple local reconstruction into a global 3d model. Thus further post-processing techniques, such as depth map filtering and fusion, were also performed to produce coherent 3D reconstruction results. Photometric and geometric consistencies were adopted as the basic criteria for depth filtering, and a visibility-based depth map fusion algorithm [16] was used to minimize depth occlusions and violations across different viewpoints and generate a coherent 3D point cloud.

Recent advances on multi-view stereo mostly follow the same pipeline as MVSNet, and make improvements on some of those procedures.

2.1. Feature Extraction

Earlier works than MVSNet have explored feature extraction using deep networks. Hartmann et al. designed multiple "siamese" CNN branches to extract features from multi-view image patches, which were determined by plane sweep depth hypotheses [17]. By aggregating features from the corresponding patches and passing through convolutional and softmax layers, the network can produce scores indicating the similarity of patches. The network was trained using a metric learning objective, which didn't directly correlated to the sub-pixel depth estimation task. Also, a point depth estimate required multiple forward passes (whose number was equal to that of depth hypotheses), thus being computationally expensive.

While MVSNet adopted 2D CNN to extract features from images, more architectures could be applied to extract more expressive multi-scale hierarchical features for matching. Im et al. extracted multi-scale feature using a spatial pyramid pooling (SPP) module [18] to gather hierarchical contextual information, similar to [19] in stereo matching. Feature Pyramid Networks (FPN) [20] were also introduced in feature extraction module to extract hierarchical information [21,22]. Chen et al. [23] further integrated instance normalization in FPN to provide more robustness to appearance changes. In D^2HC-RMVSNet, a light Dense Reception Expanded module was presented to capture multi-scale contextual information without losing resolution [24], where features maps from dilated convolutional layers with different dilated rates were concatenated and processed.

Recently, attention mechanism has been widely adopted to improve the feature expression capability. Yu et al. chose to leverage the self-attention layer to capture long-range contexts in the spatial domain [25]. By considering both the context similarity and the positional proximity [26], the self-attention layer could adaptively aggregate features across the whole image to capture important information for cost computation. To combine both low- and high-level features for MVS, Yang et al. designed a multi-level feature extraction and aggregation module [27]. A spatial path extract features that preserved affluent low-level information such as edges and corners, helping in recovering depth details. A context path, containing an attention module, introduced a large reception field to extract semantic context features. The features from these two individual path were aggregated and selected by an Squeeze-and-Excitation module [28], which produced features including both details and contexts suitable for matching. The long-range attention module was raised in [29] to enhance the image feature utilizing the long-range dependency. To be specific, the reference feature was embedded into a global descriptor through attention-based second-order pooling, such that informative features were highlighted. This global descriptor was then used to aggregate interdependence with features of both reference and source views in a soft attention manner. Such operation was leveraged across multiple levels, producing an enhanced feature pyramid for better matching.

2.2. Cost Volume Construction

A large number of works follow the variance-based method proposed by MVSNet for cost volume construction. DPSNet constructed the cost volume by directly concatenating feature maps warped from a source view and the reference view, and averaging among all pairs [18]. Though it did not provide any matching information, the costs were learned through a series of 3D convolutions, and since the reference feature provided contextual information, concatenating features improved performance over the hand-crafted distance metric (i.e., the absolute difference of the features).

Noticing that the variance-based cost volume contained redundant information and required a huge memory footprint, Inspired by [30], Xu et al. leveraged an average group-wise correlation similarity measure, operating on evenly divided channel groups between reference and source image features, to construct a lightweight cost volume [31]. Correlation operation explicitly encoded similarity, and group-wise operation maintained expression ability to some extent, so the proposed cost volume could ease the computation burden of the subsequent cost volume aggregation. Several subsequent works adopted this cost volume construction method [25,32,33].

More prior knowledge could be introduced to improve the accuracy and robustness of constructed cost volumes. Chen et al. introduced Depth-Based Attention Feature Volumes as pair-wise cost volumes in their MVSNet++ before merging them into a fused cost volume [23]. The depth mask, which indicated the reliability of the depth value of the pixels, was introduced to the cost volume construction as prior knowledge and excitation, improving the robustness and accuracy of cost volume. A Curriculum Learning framework was integrated in the construction of cost volume, making the cost volume less influenced by the depth mask during the training process. At the end of the training phase, the depth mask had no impact and thus no depth mask was needed during inference. In [34], Luo et al. claimed that the pixel-wise matching confidence volume (MCV), the cost volume based on the mean-square error of corresponding features between the reference and all source views, treated the contributions of all involved pixel pairs equally, which might be not conducive. Thus they attempted to aggregate the pixel-wise MCV into a patch-wise MCV via a learnable function. Separate functions were learned to aggregate matching information of neighboring pixels both in a single patch and among adjacent patches along the depth direction. Such enhanced MCV could highlight the importance of pixels in the reference image. Further, Luo et al. extended their patch-wise MCV to attention-enhanced matching confidence volumes [35]. All image feature maps were squeezed into channel descriptors via average pooling, and computed the variance as the

contextual channel-wise statistics of the local scene. Then it passed through a squeeze-and-excitation block to obtain attentional channel-weighted vector, which was used to enhance the cost volume. Such attention enhancement combined the photo-consistency information and the contextual cues in the matching cost computation and improved the matching robustness.

Note that images from different views leaded to heterogeneous image capture characteristics, variance-based cost volume construction which treated the contribution of different views equally could be sub-optimal. Yi et al. introduced a self-adaptive view aggregation to merge feature volumes of different views flexibly according to their contributions [36]. Specifically, two types of self-adaptive aggregation methods, point-wise and voxel-wise view selection, were introduced to the element-wise feature volume differences of all reference-source pairs, both of which utilized the attention mechanism for selecting important matching information in different views. The point-wise view aggregation leveraged a 2D weighted attention map, which encoded the various pixel-wise saliency in the spatial dimension and different depth hypothesis of each pixel shared the same weight. While voxel-based view aggregation learned a 3D weighted attention map that treated each pixel with different depth hypotheses differently. Voxel-based methods provided more flexibility to adaptive view aggregation, leading to more accuracy boost than the pixel-wise one at the cost of more parameters.

Visibility Occlusion reasoning is important for multi-view stereo since consistency-based matching would be invalid in these regions, making the prediction unreliable and reducing the reconstruction accuracy. Visibility indicates whether a 3D point is visible in given images, thus it could provide useful information about occlusion and reducing the negative impact on accuracy. Incorporating visibility into cost volume construction allows for adaptive view aggregation. While visibility information could be acquired reliably only when the precise 3D model is available, recent deep learning approaches attempted to estimate pixel-wise visibility jointly with multi-view stereo tasks. Zhang et al. firstly took pixel-wise visibility information into account [37]. The visibility between each reference-source pair was encoded as matching uncertainty, which was correlated to the entropy of matching probability. By assuming the depth followed a Laplacian distribution, the depth and uncertainty could be jointly estimated through maximizing the likelihood of the observed ground truth. The estimated per-pair uncertainty was then utilized in visibility-aware volume fusion, where values in the latent volume with large uncertainty would be attenuated. Chen et al. also introduced visibility information into the feature aggregation step in their point-based network [38]. An additional network took the reference-source feature pair and depth hypotheses as input, and output the visibility mask which was applied as the weight on variance-based cost volume construction. Xu et al. applied a 3D U-Net on each two-view cost volume for initial cost aggregation and probability volume generation, and the visibility score was computed as the maximum matching probability among depth hypotheses [32]. The visibility map, containing only scores above a preset threshold, was integrated in weighted-sum cost volume fusion. Such cost volume method was adopted also in [33]. In [29], the maximum matching probability was leveraged to construct a confidence score that modulated the matching probability, which was learned with a cross-entropy loss as in [39]. These works have shown that introducing the visibility improves accuracy and robustness in MVS networks as occlusion is modeled and learned explicitly.

2.3. Cost Volume Regularization

Cost volume regularization is of great importance for providing accurate depth and 3D reconstruction results. Thus, most works have made attempt to improve the performance of the cost aggregator over the design of MVSNet.

Huang et al. presented DeepMVS based on patch matching [40]. The concatenated patch features from the reference image and a source image were passed through an U-Net structure for intra-volume feature aggregation, which leveraged reference guided features to provide semantic information in the decoding process. To aggregate information from arbitrary multiple views, the element-wise max pooling was applied. A context network used the reference feature in DPSNet [18] to guide the refinement of each slice of the cost volume in cost aggregation. The output residual cost volume was then added back to the initial cost volume for refinement. The multi-scale information in the decoder of cost aggregation U-Net were all utilized to form the refined cost volume in MVSNet++ [23]. In [27], alongside the cost aggregation network was an autoencoder, which aimed at recovering the plane-sweep volume of neighboring views from the concatenated cost volume. This so-called Common and Private Features allowed the regularization network learn the common features between reference and neighboring views, resulting in a denoised cost volume for further aggregation. The authors in [34] argued that the plane-sweep volumes were essentially anisotropic, while other methods utilized approximated isotropic cost volumes, which could be detrimental. Thus, two kinds of anisotropic convolution were leveraged in the proposed hybrid 3d U-Net to anisotropically aggregated costs in spatial and depth directions. In [35], the attention mechanism was introduced in a hierarchical cost regularization process to adaptively aggregating multi-level regularized cost volumes. The hierarchical regularization process iteratively combine the regularized high-level and un-regularized low-level cost volumes using the proposed ray attention module (RAM), where the ray weighted map denoting the difference of cost volume between adjacent levels was leveraged to modulate the combination of adjacent levels of cost volumes.

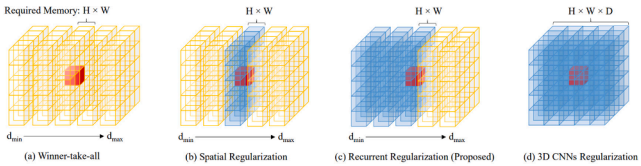

Fig. 2. Different cost volume regularization approaches [43].

Convolution networks usually performed local cost aggregation, therefore no explicit regularization was imposed on the whole structure of the depth map, especially the smoothness property. Conditional random fields (CRFs) were a kind of techniques that explicitly constraining the outputs of pixel-wise predictions, thus could be used as a post-processing step to refine the predicted depth map [40]. A more effective way was to utilize such a mechanism directly the cost aggregation stage in order to filter out the noise in the latent probability volume. Xue et al. leveraged the Conditional Random Fields (CRFs) right after 3D U-Net cost volume aggregator [21]. Setting the cost at each pixel as the unary term, the mean-field inference could be implemented as recurrent neural networks so that the network could be trained in an end-to-end manner [41]. Extending CRFs to multi-scale version enabled more global information to be incorporated for regularization. Sormann et al. proposed to integrate a belief propagation layer at the cost regularization stage, which acted as the inference module for the CRF [42]. Since the BP-layer was originally designed for label assignment problem, to make it suitable for MVS problem that commonly included multi-scale information processing and subpixel estimation, some modifications were introduced to meet the requirements of a MVS method.

Recurrent cost aggregation Due to the use of 3D convolution, the computational cost grows rapidly when it comes to a high-resolution 3D reconstruction or when the the number of input images is large, making deep MVS methods such as MVSNet fail or requiring more runtime memory. Yao et al. proposed a scalable multi-view stereo framework, R-MVSNet, which efficiently aggregated cost volume in a sequential manner [43] (see Fig. 2). Specifically, a convolutional gated recurrent unit (GRU) was applied, where the cost was aggregated both spatially using 2D convolution and in the depth direction using GRU. By stacking

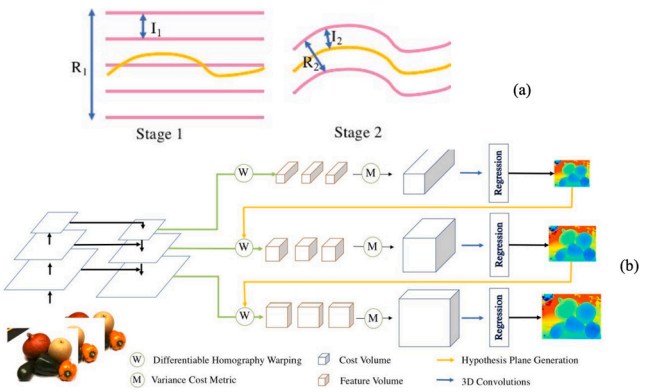

Fig. 3. (a) Depth range reduction based on coarse-to-fine strategy; (b) The architecture of CasMVSNet [22].

multiple Conv-GRU layers, the regularization results are comparable to 3D CNNs with much more efficiency in runtime memory. In D^2HC-RMVSNet, a hybrid recurrent regularization network was proposed to regularize the cost volume sequentially in depth direction [24]. A 2D U-Net was applied to each slice of cost volume, whose layer was LSTMConvCell. The proposed HU-LSTM module thus utilized rich context information to improve robustness and accuracy and was memory efficient. In order to perform larger-scale 3D reconstruction like 3D urban reconstruction from high-resolution aerial images, RED-Net chose to use 2D U-Net structure to aggregate costs at each depth dimension, and place GRU only at the end of the encoder to regularize cost sequentially along the depth direction [44]. This resulted in a significantly memory reduction. Nevertheless, such recurrent cost aggregation mechanism led to an increased running time as the depth cost slices were processed sequentially.

Coarse-to-fine architectures Considering that MVS networks are required to process 3D cost volumes, the memory and time costs grow cubically when the input resolution increases. That makes high-resolution 3D reconstruction impractical. Predicting multi-view depth maps in a coarse-to-fine manner is a promising way to effectively reduce the computational costs. The feature pyramid is obtained from either the image pyramid or Feature Pyramid Networks. First, a low-resolution coarse cost volume is constructed from the coarse level of the feature pyramid and covers the whole range of depth hypotheses. Then, the range of depth hypotheses gets narrower around the coarse depth prediction for higher-resolution fine level cost volume construction and depth prediction. The coarse prediction contains more low frequency components and the fine prediction recovers high frequency details, so the coarse-to-fine strategy can remain high accuracy while significantly reduce GPU memory usage and run-time (see Fig. 3). Several strategies for choosing the depth hypotheses for finer cost volume construction have been presented in recent works. Chen et al. tried to refine the depth

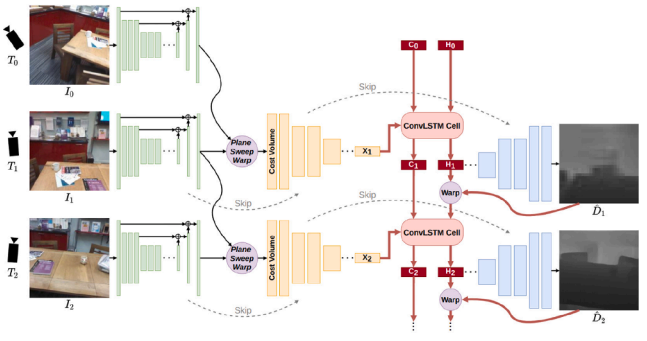

Fig. 4. Architecture of DeepVideoMVS [50] with ConvLSTM cells in cost aggregator to leverage temporal information.

predictions directly on the 3D point clouds in their PointMVSNet [45]. When the coarse depth map was generated and converted to a local point cloud, the residual between the current depth prediction and that of the ground truth was estimated iteratively. The point features for refinement were generated as the concatenation of the variance metric of features from the FPN and the 3D position of corresponding unprojected points. A sequence of point hypotheses, similar to depth hypotheses, were obtained with different displacement along the reference camera direction. Edge convolution was performed among k nearest neighbors to aggregate feature augmented point cloud locally. The augmented point features were utilized to output the probability with softmax, using which the displacement of point cloud was computed via softargmin operation. Finally the depth residual map was obtained by projecting the displacement back.

The process on 3D point cloud made PointMVSNet complicated, and later deep MVS methods focused on building the partial cost volume on the depth residual directly. Gu et al. utilized a fixed number of depth hypotheses and corresponding depth intervals at each stage in their CasMVSNet, leading to a decreasing hypothesis range around the depth prediction of the previous stage [22]. Due to its simplicity and good performance, this work has been seen as a new basis for several recent works [29]. However, a fixed depth hypothesis range would be inadequate to cover the ground truth depth value and thus unable to give a correct estimate. Cheng et al., instead, chose to use an adaptive depth hypothesis range for each pixel, whose interval length was determined by the uncertainty of the depth prediction, i.e., the variance of the probability distribution of depth hypotheses [46]. This variance-based range computation was also used in [32] for high-resolution prediction. The variance-based uncertainty, which was expected to denote the actual error, brought flexibility to determining the search range of fine refinement, leading to a higher accuracy. A similar insight was introduced in [42], where the expected 3D error in the depth dimension was directly used to determine the depth interval of fine-level depth estimation. Based on [46], Yi et al. introduced an extra range estimation module (REM) to estimate the uncertainty of the estimated depth from the probability volume via a lightweight 2D CNN [47]. For the sake of reasonable range estimation, reexamination was performed at the training stage using the new depth range and the input probability volume. A refined depth map was computed using depth candidate/probability pairs within the new range, and was forced to be close to the ground truth. Such loss strategy enabled the learned range selection to perform better than variance-based range selection. Yang et al. designed a specific criterion to set the depth hypothesis range in their CVP-MVSNet, which was determined by the corresponding pixel offset in source views [48,49,25]. Different from other coarse-to-fine methods, CVP-MVSNet was performed on the image pyramid instead of feature pyramids, and the network parameters were shared across pyramid levels. Thus inference could be performed on high-resolution images when the network was only trained on low-resolution dataset. Such coarse-to-fine strategy led to depth inference with higher compactness, accuracy and flexibility.

Temporal information in videos Multi-view stereo is quite suitable for 3D reconstruction using video streams when the camera poses are available. Such pose information could be estimated using visual-inertial odometry techniques and is readily available in standard mobile platforms (e.g., Android AR Core and iOS ARKit). Noticing that adjacent frames in videos usually have large similar observations as the camera pose varies not too much and it observes the same region from continuous varying viewpoints. On the one hand, the depth maps from video should be temporally coherent, and individual depth estimation is likely to be inconsistent. On the other hand, the overlapping regions could provide more information for accurate depth estimation. Wang et al. made the first attempt for video-based multi-view depth estimation network, namely MVDepthNet [51]. It passed the reference frame and the cost volume, which was constructed directly on multi-view images using a hand-crafted metric, to the FlowNet [52] based cost aggregation

network and regressed the multi-scale depth maps. While running in real-time, its performance was not satisfactory. Hou et al. improved the MVDepthNet structure by introducing a nonparametric Gaussian process prior with a pose-kernel structure at the bottleneck of the encoder-decoder cost aggregator [53]. By choosing suitable kernel to define "closeness" in pose-space, such prior could encourage similar poses to have resembling latent spaces in the bottleneck layer. This enabled the information fusion to use memory from previous views adaptively. Adopting the common pipeline like MVSNet, Duzceker et al. leveraged ConvLSTM module at the bottleneck of U-Net style cost volume regularizer, thus aggregating sufficient temporal information [50] (see Fig. 4). To account for the viewpoint changes, they proposed to directly warp the hidden representation based on the depth map of current view, which was also warped from the depth estimation from the previous time step. The warped hidden representation was shown beneficial for current view depth estimation. Long et al. claimed that predicting the depth maps for adjacent views jointly could promote temporal coherence and improve the depth accuracy for videos [54]. To fuse cost volumes from adjacent views, they proposed a novel Epipolar Spatio-Temporal transformer. In that module, the temporally adjacent cost volumes (including query and memory volumes) were firstly encoded as keys and values. Then the memory keys and values were warped into the query camera coordinated space. A self-attention operation was performed between query and warped memory volumes, producing an enhanced query value. Finally the query values before and after self-attention operation were fused adaptively to generate final enhance cost volume. Such module successfully propagated temporal information between frames and made the depth maps more temporally coherent. Long et al. also generated a context volume from the reference image using a 2D network, which captured the global context information. Thus, the 3D cost volume focusing on local matching information was decoupled from learning contexts. Explicitly modeling the context and matching volumes via the proposed hybrid cost regularization network could achieve faster speed and bear less computational burden than previous heavy 3D cost aggregators.

2.4. Depth Regression and Post Processing

Depth regression In most works, the depth map was estimated from the probability volume using the softargmin operation as MVSNet, and the L_1 loss or other robust regression losses were used for network training. While MVSNet uniformly sampled the hypotheses directly in the depth space, in [31] the authors chose to sample the hypotheses uniformly in the inverse depth space, making that hypotheses projected onto neighboring images could be uniformly distributed along the epipolar line. This resulted in a simpler correspondence search problem like disparity regression in stereo matching [15]. In [50] the inverse depth was linearly regressed using the sigmoid activation. For discrete depth outputs from recurrent MVS methods like [43], a variational depth map refinement based on multi-view photo-consistency could be used to reach sub-pixel accuracy.

Training loss and strategy As the probability volume explicitly presented the probability distribution over the discrete depth hypothesis, the cross-entropy loss between the predicted probability and the one-hot ground truth depth label could be used to train the network. Then the maximum depth hypothesis was selected as depth estimation. Such loss was usually used in networks with recurrent cost aggregation along depth dimension [43,44,24]. In MVSNet++ [23], more constraints were introduced for supervision, including the direct supervision on depth outputs, and the consistency between reference and warped source images at both image and feature levels. The consistency constraints could reduce the influence of occlusions, illumination and structural changes. Supervision on the spatial depth gradients was leveraged in [35] along with the depth loss function. In [55], the confidence masks were learned via an additional CNN, which mostly denoted the visibility and were introduced to the loss function to downweight uncertain pixels.

Suitable viewpoints should be determined to select the best source images for model training. In MVSNet [14], the best two neighboring views was selected according to a global view selection score. This strategy encouraged the source images to have strong visibility association with the reference image, leading to imbalance between positive and negative samples of visibility reasoning. An anti-noise training strategy that introduces disturbing views was put forward in [32], where the best and worst two neighboring views was selected according to the global view selection score. Similarity, four random views were chosen from the ten best source ones in [33], increasing the diversity at training time and improving robustness for visibility estimation.

Depth map based MVS networks trained with supervision on depth maps didn't emphasis the geometric constraint on 3D space, leading to erroneous depth prediction especially in areas with low texture. Therefore, attempts have been made to incorporate more geometric constraints into the training of MVS networks. Surface normals, which denote the local geometry of 3D scenes, are a reasonable choice. Kusupati et al. used an extra network to estimate surface normal from the feature cost volume and a world coordinate volume [56]. Accompanied with supervision on depth and normal maps, a novel depth-normal consistency loss was proposed to enforce consistency between the estimated depth and normal maps, by enforcing constraints on the spatial depth gradients in the pixel coordinate space. An combined normal map was introduced in [57], where local surface normals captured the structure of non-planar regions and average surfce normals were assigned to planar regions. Loss functions on both depth and normals were designed in a occlusion-aware manner for further refinement. With surface normals, the reconstructed scene became more structured, particularly in man-made indoor environment.

Refinement Networks Concerning the computational resource restriction, multi-view stereo networks usually produce a lower-resolution depth map than the original images, i.e., 1/4 resolution. Fine-grained details are lost at that resolution and simply using bilinear upsampling doesn't help in recovering these details. Thus an additional refinement network could be introduced to recover details in depth maps as in MVSNet. The spatial propagation network [58] was used for depth refinement in [59] under the guidance of the affinity matrix computed from the image feature. Directly upsampling depth maps to the original resolution and refine the depth map with the reference image was sufficient for [33], using the predicted depth residuals based on MSG-Net [60]. Some refinement networks produced depth maps of the original image resolution. For example, In [34], the reference feature map and the latent probability volume were concatenated and passed through the refinement network to obtained the high-resolution latent probability volume, and in [54] the initial depth maps were progressively upsampled to the full resolution using the proposed RefineNet.

Depth filtering and fusion As the depth estimation was noisy, to fuse depth maps from different views, depth filtering methods are often required to select reliable depth estimates to obtain a consistent point cloud. Two filtering criterions were introduced in [14] to discard the wrongly predicted depth values: the photometric consistency to filter out the depth value with low peak confidence (maximum probability) as the obviously untrustworthy prediction, and the geometric consistency to filter out inconsistent depth values across adjacent images whose reprojection points were far away from their locations and their depth predictions differed too much. Such criterion were adopted and developed in follow-up works [43,34]. To improve the produced depth maps for high-resolution scenes with large depth ranges, in [35], the refined process based on maximizing the multi-view photometric consistency with pixel level view selection was introduced as in [61]. A dynamic consistency checking algorithm was introduced in D^2HC-RMVSNet [24] to select reliable depth values. Photometric consistency and geometric cycle consistency were both considered, and the photometric and depth reprojection errors was computed to check the reliability of a depth value and filter out erroneous estimates. Similar photometric and geometric consistencies were utilized to select reliable depths in [36].

Besides, the authors were concerned about inaccurate depths with low confidences at higher resolution fine estimated depth map. They proposed a multi-metric pyramid depth aggregation that replaced unreliable depth estimations at the higher scale by reliable depth estimations at the lower scale. Such strategy progressively propagated to replace the mismatched depths at higher scales to resolve the matching ambiguity, improving the robustness and completeness of 3D point cloud.

2.5. Efficient Multi-view Stereo Approaches

Most learning-based MVS approaches focused on improving the 3D reconstruction quality. However, the huge computational cost required to process multiple views and produce dense depth maps has prohibited these computational expensive methods from real-world applications, especially on high-resolution images. Specifically, the memory requirement for processing a commonly used 3D cost volumes scales cubically with the resolution. This motivated several attempts for memory and computational efficient multi-view stereo approaches. Since the cost volume regularization, which consisted of several 3D convolution layers, often consumed the most memory and computational time among all stages, there have been notable attempts focusing on reducing the computation of this stage. Two of the representative ideas are recurrent cost aggregation along the depth dimension and the coarse-to-fine networks as mentioned above. The former reduces the memory usage at the cost of significantly increased inference time, while the latter can successfully reduce the memory and computational time without performance drop, thus have been a promising choice.

There are also some network designs different from the MVSNet pipeline that focuses on the efficiency of MVS reconstruction. Yu et al. presented a sparse-to-dense depth map estimation framework, Fast-MVSNet, in [62]. They constructed a sparse cost volume with a dilation rate of 2 and performed 3D cost aggregation on this sparse cost volume. Such operation leveraged larger contextual information and produced a sparse depth map with low memory costs, while losing finer details. To recover these details and obtain a dense depth map, an efficient propagation module was applied to the sparse depth map under the guidance of the reference image, acting as a generalization of the bilateral upsampler. Finally, the Gauss–Newton algorithm was chosen to refine the depth map due to its efficiency. The Gauss–Newton algorithm minimized the feature-metric reprojection error and could be implemented as a layer in a neural network, thus the network was able to learn suitable features for efficient optimization. Sinha et al. followed a similar sparse-to-dense formulation, but they constructed the sparse depth map based on the framework of keypoint detection, matching and triangulation [63]. In detail, a SuperPoint-like network [64] was used for interest point detection and description from reference and source images, distilling the output of the original SuperPoint network for detection network training. The algebraic triangulation approach [65] was leveraged to obtained 3D points from the detected interest points, where an over-determined system of equations on homogeneous 3D coordinate vector was solved via differentiable SVD. The depth was directly acquired as the z coordinate of the triangulated points, formulating a sparse depth map. which was then densified with a sparse-to-dense network under the guidance of the reference image feature. This method completely avoided cost volume construction, thus improving efficiency significantly over the cost volume based approaches. Most recently, Wang et al. borrowed the idea from the seminal Patchmatch algorithm [66] as it has seen a success in efficient traditional stereo matching benefited from the inherent spatial coherence of depth maps. They proposed PatchmatchNet, a coarse-to-fine network consisting of modules that mimicked the steps of traditional Patchmatch [33]. The proposed learnable Patchmatch consisted of three steps at each stage of the coarse-to-fine framework: 1) Initialization: generated the initial depth hypotheses for Patchmatch at the current stage, random initialization on the first stage and uniformly sampling in a decreased range around the previous estimation on other stages as in [22]; 2) Propagation: propagated hypotheses to neighbors. In Patchmatchnet propagation was performed in an adaptive manner based on extracted deep features and the Deformable Convolution Networks (DCN) implementation [67]; 3) Evaluation: computed the matching costs for all depth hypotheses and produced a softargmin prediction. An adaptive evaluation module was elaborately designed for cost aggregation, consisting of matching cost computation using plane sweep stereo, group-wise correlation and visibility-aware cost fusion [32] and adaptive spatial cost aggregation using DCN [68,69] in place of the costly 3D CNN aggregation. Since the latter two network didn't apply any 3D cost volume regularization, the memory consumption and run-time were largely saved. Patchmatch benefited from the traditional wisdom and achieved a competitive performance as the 3D cost aggregation methods.

2.6. Self-supervised Learning Approaches

Currently most deep learning based multi-view stereo methods require large-scale 3D ground truth as supervision to reach the greater performance than traditional methods. However, precise large-scale 3D ground truth data are not easy to acquire. Even with sufficiently abundant labelled training data, the generalization ability of MVS networks would be hindered in never-seen-before open-world scenarios. Thus unsupervised/self-supervised learning based MVS approaches are highly demanded. Knot et al. presented the first unsupervised multi-view stereo framework using only images from novel views as the supervisory signal [70]. To train MVSNet in an unsupervised manner, in combination with common structured similarity (SSIM) and depth smoothness objectives, they proposed a robust photometric consistency loss, enforcing photometric consistency and first-order consistency of valid pixels between the reference and warped source views, considering only top-K out of all views. Dai et al. presented a symmetric network to all views, i.e., it predicted the depth maps for all views simultaneously [59]. They designed unsupervised learning objectives based on view synthesis and cross-view consistency on both brightness and depth. Multi-view occlusion reasoning was also performed based on cross-view depth consistency check to avoid the unreliable occluded regions from participating in unsupervised training. Mallick et al. leveraged the model-agnostic meta-learning (MAML) framework [71] to learn adaptive feature representations for multi-view stereo reconstruction with view synthesis based self-supervised loss [55]. Recent unsupervised learning works claimed that the view synthesis based self-supervised learning objective, which assuming that the same point have a constant color observation among different views, is not reliable in real world where the environmental light condition varies. Thus more constraints are introduced into self-supervised objectives to address the color constancy ambiguity issue. A novel multi-metric loss function was introduced in [72], in which an feature-wise view synthesis loss function was raised based on the pretrained VGG16 network, accompanied with the common pixel-wise view synthesis loss. Such loss provided inherent constraints from different perspectives of matching correspondences. A normal-depth consistency loss was also incorporated to enforce the continuity of depth maps in 3D spaces, further improving the accuracy. Xu et al. incorporated extra priors of semantic correspondence and data augmentation consistency in self-supervised loss [73]. For the prior of semantic consistency, as the semantic labels were not available for training an extra segmentation network, the authors adopted the non-negative matrix factorization (NMF) [74] of pretrained VGG features for unsupervised co-segmentation among multi-view images. Then the semantic consistency loss could be applied between the segmentation maps of reference and warped source views. For the prior of data augmentation consistency, several data augmentation approaches (i.e. Cross-view Masking, Gamma Correction, Color Jitter and Blur) were utilized to form augmented samples, and constrained the depth outputs of original data and augmented samples to be the same as regularization. Following the coarse-to-fine architecture [48], Yang et al. proposed a

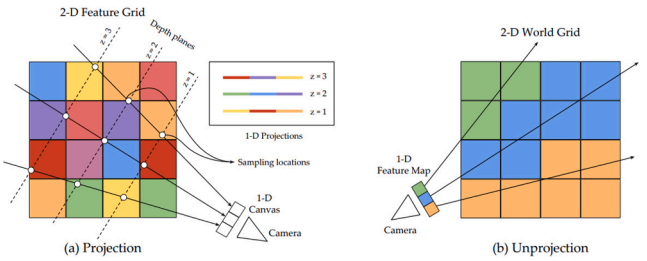

Fig. 5. Feature projection and unprojection operation for volumetric based methods [78].

iterative self-training strategy to trained the network in a self-supervised way [75]. After initializing the network with view synthesis unsupervised loss, the depth maps were estimated and filtered to produce pseudo labels as supervision for further network training. To ensure the reliability of pseudo labels, the low-resolution depth maps were refined on higher resolution training images which carried more discriminative features, thanks to the coarse-to-fine architecture and shared network parameters across scales. The high-resolution refined depth were filtered via cross-view depth consistency check and visibility check, and then fused as a pseudo mesh using Screened Poisson Surface Reconstruction method [76]. The refined pseudo labels were rendered from the mesh and used as supervision in self-training. These recent progresses improved the performance of unsupervised MVS networks and made multi-view stereo reconstruction in real world promising.

3. Volumetric based methods

Depth maps only present 3D reconstructions locally in the reference view. The globally consistent 3D models, e.g., global point clouds, are thus constructed indirectly. Volumetric representations, based on volume occupancy in 3D space, could present the scene geometry in whole and recently several works have tried to develop deep networks that directly reconstruct consistent and efficient voxel-based 3D models.

Early attempts focused on the reconstruction of object geometry. SurfaceNet presented a network that reconstructed a 2D surface from the voxelized 3D space [77]. It predicted a binary label for each voxel whether it was on the surface or not. First of all, each input image was unprojected to the 3D voxel space along the ray from the viewpoint, and each voxel stored the corresponding pixel value. This process produced the color voxel cube (CVC) representation for each view, and then the pair of CVCs from any two views were sent into a network that predicted the on-surface probability for each voxel position. This surface prediction network was training using a class-balanced cross-entropy loss. For multi-view stereo surface prediction, a fusion approach of surface probability predictions from all view pairs was needed. Taking into consideration the factors like viewpoint differences and occlusions, a triplet network was trained to produce the relative weights for each view pair, which were later used in the weight average of predicted surface probabilities for fusion. Kar et al. leveraged the 3D geometry through the projection and unprojection of image features along viewing rays [78] (see Fig. 5). The proposed network utilized the discrete grid as internal representation of 3D world and unprojected the image features into the 3D feature grids by rasterizing the viewing rays. Such operation aligned the features along the epipolar lines. A recurrent neural network was utilized to process the feature grids sequentially to produce the cost grids. This process could be regarded as an implicit local matching procedure across views. As the local matching cost grids were typically noisy, a 3D UNet were applied to smooth the cost grid taking context into account. Based on the cost grids, either a 3D volumetric occupancy map or projected depth maps of all views could be generated, denoting the 3D geometry of models. These two early works, however, were limited to process 3D models at the object level.

As the voxel representation covers the whole 3D space, the number of voxels grows cubically when the 3D model gets larger. This significantly influences the computational requirement in both time and memory and impedes the voxel representation learning in scene reconstruction. Murez et al. presented the first multi-view stereo networks for indoor scene reconstruction that directly regressed to 3D [79]. In their work, the Truncated Signed Distance Function (TSDF) representation was introduced to represented the 3D scenes effectively. The input image sequences were passed through a 2D CNN backbone to extract features, which were then back projected into the 3D space to form a 3D voxel volume and accumulated using a weighted running average similar to TSDF fusion. A 3D convolutional UNet was utilized to refine the 3D voxel volume and predict the TSDF values. While being able to recover a complete 3D scene structure, the voxel grid of the whole scene was processed with 3D convolutions, which prohibited this network from processing a relatively large-scale scene and running efficiently in real-time.

Most recently, Sun et al. proposed a framework, NeuralRecon, for real-time 3D scene reconstruction from videos [80]. Adopting the sparse TSDF volumes as the scene representation as in [79], several improvements were presented in this work to deal with larger scale scenes and run in real-time. NeuralRecon incrementally reconstructed the local geometry in a global view-independent 3D volume with a set of key frames in a local fragment. The images in a local fragment were passed through the feature pyramid network to extract multi-scale features, and back projected them into 3D feature volumes along viewing rays. These feature volumes from different views were integrated into a single feature volume via averaging according to the visibility weights, and a 3D sparse convolution was applied to extract 3D geometric features. To refine geometric feature volumes and make the reconstruction consistent across all local fragments, a 3D convolutional variant of Gated Recurrent Unit (GRU) module was used, fusing the 3D geometric feature with the hidden state at each timestamp. The updated hidden state was passed through the MLP layers to predicted the TSDF volumes. In order to efficiently process the feature volumes in a large-scale scene, the coarse-to-fine strategy was introduced in NeuralRecon to gradually refine the TSDF volumes at each level when the density of sparse voxels increased gradually. At each level, the TSDF volumes contained the occupancy score, the confidence of a voxel being within a given TSDF truncation distance, and the SDF value. The occupancy loss with the binary cross-entropy of occupancy and the SDF loss with the L_1 distance were applied for training. Božič et al. adopted a similar coarse-to-fine architecture for volumetric based scene reconstruction, while they attempted to utilize the transformer architecture to fuse the unprojected 3D image features from each view into a global 3D feature [81]. The transformer network could attentively leverage the most informative features for scene reconstruction.

4. Datasets and Performance Evaluation

In this section, several widely used datasets for multi-view stereo evaluation are introduced. Also, the experimental results of several representative multi-view stereo network on two mainstream datasets, DTU as well as Tanks and Temples, are summarized.

4.1. DTU

DTU [82] is a large-scale dataset containing 128 scenes in a controlled laboratory environment, whose reference models are captured using a structured light scanner. Each scene is scanned at the same 49 or 64 camera positions under 7 different lighting conditions, producing RGB images whose resolution is 1200×1600 pixels. As the dataset covers a variety of objects and materials, it is well-suited to train and test deep learning MVS methods under realistic conditions. As reference model is a point cloud, the ground truth depth maps should be rendered from the mesh models, which are generated from the reference model using a surface reconstruction method, e.g. the Screened Poisson

Surface Reconstruction method [76]. In most works, the dataset is divided into three subsets, i.e., the training, testing and validation split following [77].

Metrics The evaluation protocol takes the scanned reference models and an MVS reconstruction result, and computes the (mean and the median) point-wise reconstruction error, which quantifies the fitness of MVS reconstruction to the scanned model. Specifically, accuracy and completeness are used as evaluation measures. Accuracy (precision) measures the mean or median absolute distance of each reconstructed point to the nearest point in the ground truth reference model. Precision (recall) measures the mean or median absolute distance of each ground truth point to the nearest point in the reconstructed point cloud. In some works, the accuracy and completeness are computed as the percentage of points whose error with respect to the corresponding distance is smaller than a preset threshold. Usually, the overall score is also measured as the average of mean accuracy and mean completeness, denoting the overall quality of the reconstruction results.

4.2. Tanks and Temples

Tanks and Temples [83] is a large-scale benchmark in realistic conditions, whose ground truth data is captured using an industrial laser scanner. This dataset contains both indoor environments and outdoor scenes with small depth ranges. High-resolution videos are contained in this dataset, which support the evaluation of video based multi-view stereo methods. All scenes in this dataset are divided into intermediate and advanced groups according to the difficulty caused by due to the scale, complexity, and other complicating factors. It also provides an online benchmark for performance evaluation.

Metrics The F-score at a given threshold is the metric for performance evaluation on the Tanks and Temples dataset. It is an integrated measure computed as the harmonic mean of accuracy and precision percentages at the given threshold. Comparing to the arithmetic mean, the harmonic mean is more sensitive to small values and can reflect the bad performance when either accuracy or completeness is of a low value.

4.3. ETH3D

ETH3D [84] is a multi-view stereo dataset containing a variety of indoor and outdoor scenes captured by both high-resolution DSLR imagery and synchronized low-resolution stereo videos. The ground truth point clouds are obtained from a high-resolution laser scanner and aligned with images using a robust photometric consistency based optimization approach. This dataset also provides an online benchmark for evaluation.

Metrics The accuracy and completeness mentioned above are leveraged to measure the consistency between the reconstructed 3D models and the laser point clouds. Both metrics are evaluated in the distance range from 1 cm to 50 cm. Since the definition of accuracy and completeness are susceptible to the density of both the reconstructed and the ground truth point clouds, the 3D space is discretized into voxels with small size length, and then both metrics are evaluated over all voxels. In order to obtain a single measure for ranking the results, the F-score is also computed.

4.4. ScanNet

ScanNet [85] is an RGB-D video dataset containing more than 2.5 million frames from 1500 indoor scenes with ground truth 3D camera poses, automated surface reconstructions and crowd-sourced semantic annotation. As it provides a large amount of videos and corresponding 3D surface models, this dataset has been used for training multi-view stereo networks, especially for those with video inputs, though it is not designed for this task. Both 2D depth metrics and 3D metrics are leveraged for benchmarking the 3D reconstruction. The accuracy, completeness and F-score mentioned above are commonly used 3D metrics. For 2D depth map evaluation, various metrics on monocular depth estimation could be used. A detailed list of metrics is presented in [79].

Table 1
Multi-view stereo evaluation of the different methods on DTU dataset (all metrics are computed as the mean, lower is better).

Method	Accuracy (mm)	Completeness (mm)	Overall Score (mm)
MVSNet [14]	0.396	0.527	0.462
R-MVSNet [43]	0.383	0.452	0.417
P-MVSNet [34]	0.396	0.527	0.462
MVSCRF [21]	0.371	0.426	0.398
Point-MVSNet [45]	0.343	0.411	0.376
VA-Point-MVSNet [38]	0.359	0.358	0.359
MVS^{2} † [59]	0.760	0.515	0.637
CIDER [31]	0.417	0.437	0.427
CasMVSNet [22]	0.325	0.385	0.355
UCSNet [46]	0.338	0.349	0.344
CVP-MVSNet [48]	0.296	0.406	0.351
Att-MVS [35]	0.383	0.329	0.356
Fast-MVSNet [62]	0.336	0.403	0.370
D^{2}HC-RMVSNet [24]	0.395	0.378	0.386
MVSNet++ [23]	0.407	0.345	0.376
Vis-MVSNet [37]	0.369	0.361	0.365
Meta_MVS† [55]	0.594	0.779	0.687
PVSNet [32]	0.337	0.315	0.326
BP-MVSNet [42]	0.333	0.320	0.327
JDACS† [73]	0.571	0.515	0.543
JDACS-MS† [73]	0.398	0.318	0.358
PatchmatchNet [33]	0.427	0.277	0.352
Self-supervised-CVP-MVSNet† [75]	0.308	0.418	0.363
LANet [29]	0.320	0.349	0.335
AACVP-MVSNet [25]	0.357	0.326	0.341
DDR-Net [47]	0.339	0.320	0.329
M^{3}VSNet† [72]	0.636	0.531	0.583

4.5. BlendedMVS

BlendedMVS [86] is a recently proposed large-scale multi-view stereo dataset, aiming at providing sufficient training ground truth for deep learning based methods. Since large-scale ground-truth models are expensive to be scanned, this dataset is synthetic and the training images and depth maps are rendered from textured 3D mesh models. A low-cost data generation pipeline with a novel fusion method, introducing varying lightings, was proposed to generate training ground truth automatically. This dataset contains 113 well selected 3D textured models, which covers different scenes, such as architectures, street-views, sculptures and small objects. Each scene includes 20 to 1,000 images, and a total of 17,818 images rendered along unstructured camera trajectories are in this dataset. All training images and ground truth depth maps are unified into a resolution of 1536×2048. This larger-scale, highly accurate synthetic dataset can greatly improve the generalization ability of the trained models.

Metrics For quantitative evaluation, both the 2D depth map validation and 3D point cloud evaluation metrics are utilized. 2D depth map validation considers the following three measures: 1) the end point error (EPE) using the average L_1 loss; 2) the >1 pixel error defined as the ratio of pixels whose L_1 error is larger than the given threshold of 1 pixel; and 3) the >3 pixel error. The point cloud evaluation metrics are accuracy, completeness and F-score as mentioned above.

4.6. Beihang Aerial 3D dataset

Beihang Aerial 3D dataset is a large-scale benchmark for aerial images based 3D reconstruction. The dataset contains 10 outdoor scenes in both cities and suburbs. The DJI drone captures over 2000 images at an

Table 2
Multi-view stereo evaluation of the different methods on Tanks and Temples dataset (F-scores of each scene and the mean F-score, higher is better).

Method	Mean	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train
MVSNet [14]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.9	34.69
P-MVSNet [34]	55.62	70.04	44.64	40.22	65.2	55.08	55.17	60.37	54.29
Point-MVSNet [45]	48.27	61.79	41.15	34.2	50.79	51.97	50.85	52.38	43.06
VA-Point-MVSNet [38]	48.7	61.95	43.73	34.45	50.01	52.67	49.71	52.29	44.75
CasMVSNet [22]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51
UCSNet [46]	54.83	76.09	53.16	43.03	54	55.6	51.49	57.38	47.89
CVP-MVSNet [48]	54.03	76.5	47.74	36.34	55.12	57.28	54.28	57.43	47.54
Att-MVS [35]	60.05	73.9	62.58	44.08	64.88	56.08	59.39	63.42	56.06
Fast-MVSNet [62]	47.39	65.18	39.59	34.98	47.81	49.16	46.2	53.27	42.91
D^2HC-RMVSNet [24]	59.2	74.69	56.04	49.42	60.08	59.81	59.61	60.04	53.92
MVSNet++ [23]	49.12	62.64	38.49	39.6	48.4	54.95	51.69	52.28	44.92
Vis-MVSNet [37]	60.03	77.4	60.23	47.07	63.44	62.21	57.28	60.54	52.07
BP-MVSNet [42]	57.6	77.31	60.9	47.89	58.26	56	51.54	58.47	50.41
JDACS† [73]	45.48	66.62	38.25	36.11	46.12	46.66	45.25	47.69	37.16
Self-supervised-CVP-MVSNet† [75]	46.71	64.95	38.79	24.98	49.73	52.57	51.53	50.66	40.45
LANet [29]	55.7	76.24	54.32	49.85	54.03	56.08	50.82	53.71	50.57
AACVP-MVSNet [25]	58.39	78.71	57.85	50.34	52.76	59.73	54.81	57.98	54.94
DDR-Net [47]	54.91	76.18	53.36	43.43	55.2	55.57	52.28	56.04	47.17
M^3VSNet† [72]	37.67	47.74	24.38	18.74	44.42	43.45	44.95	47.39	30.31

altitude of 150 meters. Each image is of 3648 × 5472 spatial resolution and has an exact pose calculated using GPS and RTK, which can be used for multi-view stereo and 3D reconstruction. The ground truth is reconstructed 3d models represented by point clouds and meshes. This dataset will be available soon.

4.7. Performance

We presented the performance of most representative approaches on two mainstream datasets, DTU and Tanks and Temples. The performance summary of each method on DTU is shown in Table 1. The performance summary of each method on the Intermediate group datasets of Tanks and Temples is shown in Table 2. † denotes self-supervised methods that trained without annotated 3D model ground truth. It is worth mentioning that these two datasets are captured with unstructured viewpoints which means methods utilizing video inputs, such as [79,50,54,80], can't be evaluated on these two datasets.

On the DTU datasets, top methods [32,42,47,46,48,33] all adopted the coarse-to-fine architectures, and adaptive aggregation [42,33,35] as well as visibility reasoning [32,33] may also impact the overall accuracy. PatchmatchNet, which was designed for computational efficiency, have achieved the best completeness performance and competitive overall performance, thus showing values in practical use. And self-supervised methods have reached a comparable level in both accuracy [75] and completeness [73]. On the Tanks and Temples datasets, methods with a coarse-to-fine architectures consistently boost the overall performance, e.g. [22,37,25,42], and methods with adaptive cost aggregation such as [35,42] also have remarkable results. Different from DTU dataset, however, self-supervised approaches still fall behind in overall performance and more attempts are needed.

5. Summary and Future Directions

In this paper, we reviewed the most recent deep learning methods for multi-view stereo tasks. Both depth map based and volumetric based methods are covered, and the former are presented in detail. We point out several aspects of the network design that lead to the performance improvement of depth map based networks, such as visibility analysis, recurrent cost aggregation, coarse-to-fine-strategy and temporal information fusion. The performance of several approaches were summarized.

While deep learning based methods have witnessed a huge progress, several challenges still remain and thus hamper the networks from real world applications.

- One of the greatest concern is the network complexity. 3D convolution have brought large improvement on cost volume regularization, but at a cost of high computational time and runtime memory requirement. Though several attempts has made to reduce the runtime memory (e.g. recurrent cost aggregation) and the time usage during inference (e.g., coarse-to-fine architectures), it is still challenging to produce high-quality 3D reconstruction from *a long video sequence*, *high-resolution images* and at a *larger scale*. For example, 3D scene reconstruction from aerial vehicles is in high demand and faces the above challenges to generate fine-grained city-scale reconstructions, and scene reconstruction from high-resolution videos captured by smartphones is also in need of efficient and real-time MVS approaches.
- The gap between depth maps and point clouds still exists and a unified framework for depth map based coherent scene reconstruction is welcomed. Another coherent reconstruction approach, the volumetric based method, have made promising progresses but still facing the complexity for reconstructing large-scale scenes, especially complicated outdoor scenes. The potential for mesh representation based deep learning approaches is also worth analyzing [87].
- Another concern comes from the self-supervised training and the domain adaption techniques for multi-view stereo, as the labeled training data from a wide range of varying scenes are quite difficult to obtain, and the performance of self-supervised multi-view stereo methods is still not satisfactory.
- The number of available datasets and the diversity of data are not adequate, such as high-resolution (4 K) datasets, large-scale aerial MVS datasets and video datasets for MVS network training.

Apart from the above challenges, several recent advances in machine learning and computer vision have potentials for improving the performance, such as Transformer [81] or contrastive learning. It is worth mentioning that the recent Neural radiance fields (NeRF) [88] and its following works have made impressive results on neural rendering by encoding the scene structure implicitly in networks. Thus how to leverage such implicit representation of scenes for multi-view stereo tasks would be an interesting topic [89].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (4202039) and National Natural Science Foundation of China project No. 61772057.

References

[1] C. Yildirim, Cybersickness during vr gaming undermines game enjoyment: A mediation model, Displays 59 (2019) 35–43.
[2] H. Kang, J. Ko, H. Park, H. Hong, Effect of outside view on attentiveness in using see-through type augmented reality device, Displays 57 (2019) 1–6.
[3] M. Emoto, Depth perception and induced accommodation responses while watching high spatial resolution two-dimensional tv images, Displays 60 (2019) 24–29.
[4] Z. Gao, G. Zhai, H. Deng, X. Yang, Extended geometric models for stereoscopic 3d with vertical screen disparity, Displays 65 (2020) 101972.
[5] N. Sugita, K. Sasaki, M. Yoshizawa, K. Ichiji, M. Abe, N. Homma, T. Yambe, Effect of viewing a three-dimensional movie with vertical parallax, Displays 58 (2019) 20–26.
[6] B. Lu, Y. He, H. Wang, Stereo disparity optimization with depth change constraint based on a continuous video, Displays (2021) 102070.
[7] B. Lu, L. Sun, L. Yu, X. Dong, An improved graph cut algorithm in stereo matching, Displays 69 (2021) 102052.
[8] C. Yan, G. Pang, X. Bai, C. Liu, N. Xin, L. Gu, J. Zhou, Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss, IEEE Trans. Multimedia (2021).
[9] C. Wang, X. Wang, X. Bai, Y. Liu, J. Zhou, Self-supervised deep homography estimation with invertibility constraints, Pattern Recogn. Lett. 128 (2019) 355–360.
[10] X. Ning, P. Duan, W. Li, S. Zhang, Real-time 3d face alignment using an encoder-decoder network with an efficient deconvolution layer, IEEE Signal Process. Lett. 27 (2020) 1944–1948.
[11] S. Qi, X. Ning, G. Yang, L. Zhang, P. Long, W. Cai, W. Li, Review of multi-view 3d object recognition methods based on deep learning, Displays (2021) 102053.
[12] W. Cai, D. Liu, X. Ning, C. Wang, G. Xie, Voxel-based three-view hybrid parallel network for 3d object classification, Displays (2021) 102076.
[13] C. Wang, X. Bai, X. Wang, X. Liu, J. Zhou, X. Wu, H. Li, D. Tao, Self-supervised multiscale adversarial regression network for stereo disparity estimation, IEEE Transactions on Cybernetics (2020).
[14] Y. Yao, Z. Luo, S. Li, T. Fang, L. Quan, Mvsnet: Depth inference for unstructured multi-view stereo, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 767–783.
[15] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, A. Bry, End-to-end learning of geometry and context for deep stereo regression, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 66–75.
[16] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.M. Frahm, R. Yang, D. Nistér, M. Pollefeys, Real-time visibility-based fusion of depth maps, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.
[17] W. Hartmann, S. Galliani, M. Havlena, L. Van Gool, K. Schindler, Learned multi-patch similarity, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1586–1594.
[18] S. Im, H.G. Jeon, S. Lin, I.S. Kweon, Dpsnet: End-to-end deep plane sweep stereo, in: International Conference on Learning Representations, 2019.
[19] J.R. Chang, Y.S. Chen, Pyramid stereo matching network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5410–5418.
[20] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
[21] Y. Xue, J. Chen, W. Wan, Y. Huang, C. Yu, T. Li, J. Bao, Mvscrf: Learning multi-view stereo with conditional random fields, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4312–4321.
[22] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, P. Tan, Cascade cost volume for high-resolution multi-view stereo and stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2495–2504.
[23] P.H. Chen, H.C. Yang, K.W. Chen, Y.S. Chen, Mvsnet++: Learning depth-based attention pyramid features for multi-view stereo, IEEE Trans. Image Process. 29 (2020) 7261–7273.
[24] J. Yan, Z. Wei, H. Yi, M. Ding, R. Zhang, Y. Chen, G. Wang, Y.W. Tai, Dense hybrid recurrent multi-view stereo net with dynamic consistency checking, in: European Conference on Computer Vision, Springer, 2020, pp. 674–689.
[25] A. Yu, W. Guo, B. Liu, X. Chen, X. Wang, X. Cao, B. Jiang, Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction, ISPRS Journal of Photogrammetry and Remote Sensing 175 (2021) 448–460.
[26] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, in: Advances in Neural Information Processing Systems, 2019, pp. 68–80.
[27] H.C. Yang, P.H. Chen, K.W. Chen, C.Y. Lee, Y.S. Chen, Fade: Feature aggregation for depth estimation with multi-view stereo, IEEE Trans. Image Process. 29 (2020) 6590–6600.
[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[29] X. Zhang, Y. Hu, H. Wang, X. Cao, B. Zhang, Long-range attention network for multi-view stereo, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3782–3791.
[30] X. Guo, K. Yang, W. Yang, X. Wang, H. Li, Group-wise correlation stereo network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3273–3282.
[31] Q. Xu, W. Tao, Learning inverse depth regression for multi-view stereo with correlation cost volume, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020a, pp. 12508–12515.
[32] Q. Xu, W. Tao, Pvsnet: Pixelwise visibility-aware multi-view stereo network, arXiv preprint arXiv:2007.07714 (2020b).
[33] F. Wang, S. Galliani, C. Vogel, P. Speciale, M. Pollefeys, Patchmatchnet: Learned multi-view patchmatch stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14194–14203.
[34] K. Luo, T. Guan, L. Ju, H. Huang, Y. Luo, P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10452–10461.
[35] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, Y. Luo, Attention-aware multi-view stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1590–1599.
[36] H. Yi, Z. Wei, M. Ding, R. Zhang, Y. Chen, G. Wang, Y.W. Tai, Pyramid multi-view stereo net with self-adaptive view aggregation, in: European Conference on Computer Vision, Springer, 2020, pp. 766–782.
[37] J. Zhang, Y. Yao, S. Li, Z. Luo, T. Fang, Visibility-aware multi-view stereo network, British Machine Vision Conference (2020).
[38] R. Chen, S. Han, J. Xu, et al., Visibility-aware point-based multi-view stereo network, IEEE transactions on pattern analysis and machine intelligence (2020).
[39] Y. Zhang, Y. Chen, X. Bai, S. Yu, K. Yu, Z. Li, K. Yang, Adaptive unimodal cost volume filtering for deep stereo matching, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 12926–12934.
[40] P.H. Huang, K. Matzen, J. Kopf, N. Ahuja, J.B. Huang, Deepmvs: Learning multi-view stereopsis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2821–2830.
[41] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1529–1537.
[42] C. Sormann, P. Knöbelreiter, A. Kuhn, M. Rossi, T. Pock, F. Fraundorfer, Bp-mvsnet: Belief-propagation-layers for multi-view-stereo, in: 2020 International Conference on 3D Vision (3DV), IEEE, 2020, pp. 394–403.
[43] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, L. Quan, Recurrent mvsnet for high-resolution multi-view stereo depth inference, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5525–5534.
[44] J. Liu, S. Ji, A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6050–6059.
[45] R. Chen, S. Han, J. Xu, H. Su, Point-based multi-view stereo network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1538–1547.
[46] S. Cheng, Z. Xu, S. Zhu, Z. Li, L.E. Li, R. Ramamoorthi, H. Su, Deep stereo using adaptive thin volume representation with uncertainty awareness, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2524–2534.
[47] P. Yi, S. Tang, J. Yao, Ddr-net: Learning multi-stage multi-view stereo with dynamic depth range, arXiv preprint arXiv:2103.14275 (2021).
[48] J. Yang, W. Mao, J.M. Alvarez, M. Liu, Cost volume pyramid based depth inference for multi-view stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4877–4886.
[49] J. Yang, W. Mao, J. Alvarez, M. Liu, Cost volume pyramid based depth inference for multi-view stereo, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[50] A. Duzceker, S. Galliani, C. Vogel, P. Speciale, M. Dusmanu, M. Pollefeys, Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15324–15333.
[51] K. Wang, S. Shen, Mvdepthnet: Real-time multiview depth estimation neural network, in: 2018 International conference on 3d vision (3DV), IEEE, 2018, pp. 248–257.
[52] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2758–2766.
[53] Y. Hou, J. Kannala, A. Solin, Multi-view stereo by temporal nonparametric fusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2651–2660.
[54] X. Long, L. Liu, W. Li, C. Theobalt, W. Wang, Multi-view depth estimation using epipolar spatio-temporal networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8258–8267.
[55] A. Mallick, J. Stückler, H. Lensch, Learning to adapt multi-view stereo by self-supervision, in: British Machine Vision Conference, 2020.
[56] U. Kusupati, S. Cheng, R. Chen, H. Su, Normal assisted stereo depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2189–2199.

[57] X. Long, L. Liu, C. Theobalt, W. Wang, Occlusion-aware depth estimation with adaptive normal constraints, in: European Conference on Computer Vision, Springer, 2020, pp. 640–657.
[58] S. Liu, S. De Mello, J. Gu, G. Zhong, M.H. Yang, J. Kautz, Learning affinity via spatial propagation networks, in: Advances in Neural Information Processing Systems, 2017, pp. 1519–1529.
[59] Y. Dai, Z. Zhu, Z. Rao, B. Li, Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry, in: 2019 International Conference on 3D Vision (3DV), IEEE, 2019, pp. 1–8.
[60] T.W. Hui, C.C. Loy, X. Tang, Depth map super-resolution by deep multi-scale guidance, in: European conference on computer vision, Springer, 2016, pp. 353–369.
[61] E. Zheng, E. Dunn, V. Jojic, J.M. Frahm, Patchmatch based joint view selection and depthmap estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1510–1517.
[62] Z. Yu, S. Gao, Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1949–1958.
[63] A. Sinha, Z. Murez, J. Bartolozzi, V. Badrinarayanan, A. Rabinovich, Deltas: Depth estimation by learning triangulation and densification of sparse points, in: European Conference on Computer Vision, Springer, 2020, pp. 104–121.
[64] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 224–236.
[65] K. Iskakov, E. Burkov, V. Lempitsky, Y. Malkov, Learnable triangulation of human pose, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7718–7727.
[66] C. Barnes, E. Shechtman, A. Finkelstein, D.B. Goldman, Patchmatch: A randomized correspondence algorithm for structural image editing, ACM Trans. Graph. 28 (2009) 24.
[67] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.
[68] H. Xu, J. Zhang, Aanet: Adaptive aggregation network for efficient stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1959–1968.
[69] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308–9316.
[70] T. Khot, S. Agrawal, S. Tulsiani, C. Mertz, S. Lucey, M. Hebert, Learning unsupervised multi-view stereopsis via robust photometric consistency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[71] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1126–1135.
[72] B. Huang, H. Yi, C. Huang, Y. He, J. Liu, X. Liu, M 3vsnet: Unsupervised multi-metric multi-view stereo network, arXiv preprint arXiv:2004.09722 (2020).
[73] H. Xu, Z. Zhou, Y. Qiao, W. Kang, Q. Wu, Self-supervised multi-view stereo via effective co-segmentation and data-augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 2, 2021, p. 6.
[74] E. Collins, R. Achanta, S. Susstrunk, Deep feature factorization for concept discovery, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 336–352.
[75] J. Yang, J.M. Alvarez, M. Liu, Self-supervised learning of depth inference for multi-view stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7526–7534.
[76] M. Kazhdan, H. Hoppe, Screened poisson surface reconstruction, ACM Transactions on Graphics (ToG) 32 (2013) 1–13.
[77] M. Ji, J. Gall, H. Zheng, Y. Liu, L. Fang, Surfacenet: An end-to-end 3d neural network for multiview stereopsis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2307–2315.
[78] A. Kar, C. Häne, J. Malik, Learning a multi-view stereo machine, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 364–375.
[79] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, A. Rabinovich, Atlas: End-to-end 3d scene reconstruction from posed images, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, Springer, 2020, pp. 414–431.
[80] J. Sun, Y. Xie, L. Chen, X. Zhou, H. Bao, Neuralrecon: Real-time coherent 3d reconstruction from monocular video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15598–15607.
[81] A. Božič, P. Palafox, J. Thies, A. Dai, M. Nießner, Transformerfusion: Monocular rgb scene reconstruction using transformers, arXiv preprint arXiv:2107.02191 (2021).
[82] H. Aanæs, R.R. Jensen, G. Vogiatzis, E. Tola, A.B. Dahl, Large-scale data for multiple-view stereopsis, Int. J. Comput. Vision 120 (2016) 153–168.
[83] A. Knapitsch, J. Park, Q.Y. Zhou, V. Koltun, Tanks and temples: Benchmarking large-scale scene reconstruction, ACM Transactions on Graphics (ToG) 36 (2017) 1–13.
[84] T. Schops, J.L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, A. Geiger, A multi-view stereo benchmark with high-resolution images and multi-camera videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3260–3269.
[85] A. Dai, A.X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5828–5839.
[86] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, L. Quan, Blendedmvs: A large-scale dataset for generalized multi-view stereo networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1790–1799.
[87] R. Shrestha, Z. Fan, Q. Su, Z. Dai, S. Zhu, P. Tan, Meshmvs: Multi-view stereo guided mesh reconstruction, arXiv preprint arXiv:2010.08682 (2020).
[88] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, in: European conference on computer vision, Springer, 2020, pp. 405–421.
[89] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, H. Su, Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo, arXiv preprint arXiv:2103.15595 (2021).